REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
30-07-2014	Final Report	1-Sep-2009 - 30-Sep-2013

4. TITLE AND SUBTITLE
Final Report: Visual Information Theory and Visual Representations for Achieving Provable Bounds in Vision-Based Control and Decision

5a. CONTRACT NUMBER
W911NF-09-1-0449

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Stefano Soatto

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of California - Los Angeles
11000 Kinross Avenue, Suite 211

Los Angeles, CA 90095 -1406

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
56765-CS.16

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
This project pursued the development of representations of visual data suitable for control and decision tasks. The fundamental premise is that traditional notions of information developed in support of communication engineering – where the task is reproduction of the source data, and nuisance factors can be easily characterized statistically – are unsuited to visual inference, where the task is decision or control, and the data formation process include scaling (that makes the continuous limit relevant) and occlusion (that makes control relevant).
Specifically, the task (or classes of task) inform what portion of the data is "informative" and what is "nuisance

15. SUBJECT TERMS
Sensing, Control, Information, Visual Representations

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Stefano Soatto
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 310-825-4840

Report Title

Final Report: Visual Information Theory and Visual Representations for Achieving Provable Bounds in Vision-Based Control and Decision

ABSTRACT

This project pursued the development of representations of visual data suitable for control and decision tasks. The fundamental premise is that traditional notions of information developed in support of communication engineering – where the task is reproduction of the source data, and nuisance factors can be easily characterized statistically – are unsuited to visual inference, where the task is decision or control, and the data formation process include scaling (that makes the continuous limit relevant) and occlusion (that makes control relevant). Specifically, the task (or classes of task) inform what portion of the data is "informative" and what is "nuisance variability." One of the peculiarities of visual processing is that most of the complexity in visual data can be ascribed to nuisance factors that affect the data but are irrelevant to the task. This notion has been made precise in [16], preceding the commencement of this project, where it was shown that the quotient of the (infinite- dimensional) set of image modulo changes of viewpoint and illumination is supported on a set of measure zero of the image domain. So, for any task that requires invariance to viewpoint and illumination (such as object detection, localization, recognition, categorization), a zero- measure set contains as much "information" as the original data. In other words, information is a thin set of visual data, for decision and control tasks.
The development of a theory of information in support of decision and control tasks, specific to visual data, is a long-term goal that has been under development since 2007, and will continue for the foreseeable future. During the course of this project, significant progress has been registered in a number of areas critical to such a development, which is described below.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
08/26/2012	9.00	Alper Ayvaci, Michalis Raptis, Stefano Soatto. Sparse Occlusion Detection with Optical Flow, International Journal of Computer Vision, (10 2011): 0. doi: 10.1007/s11263-011-0490-7
08/26/2012	10.00	E. S. Jones, S. Soatto. Visual-inertial navigation, mapping and localization: A scalable real-time causal approach, The International Journal of Robotics Research, (01 2011): 0. doi: 10.1177/0278364910388963
TOTAL:		**2**

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received		Paper
TOTAL:		

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received		Paper
07/27/2014	13.00	V. Karasev, A. Ravichandran, and S. Soatto.. Active frame, location, and detector selection for automated and manual video annotation., Proc. of the IEEE Intl. Conf. on Comp. Vis. and Patt. Recog. (CVPR). 07-JAN-14, . : ,
07/27/2014	14.00	B. Taylor, A. Ayvaci, A. Ravichandran, and S. Soatto.. Semantic video segmentation from occlusion relations via convex optimization., Proc. of the EMMCVPR (energy minimization in computer vision and pattern recognition). 01-AUG-13, . : ,
07/27/2014	15.00	K. Tsotsos, A. Pretto, and S. Soatto. . Visual-inertial ego-motion estimation for humanoid platforms., Proc. IEEE Intl. Conf. Humanoid Robots. 12-JAN-12, . : ,
08/13/2011	1.00	Alper Ayvaci, Michalis Raptis, Stefano Soatto. Optical flow and occlusion detection with convex optimization, Proceedings of the Neuro Information Processig Systems (NIPS). 01-DEC-10, . : ,
08/13/2011	5.00	Stefano Soatto, Michalis Raptis. Tracklet Descriptorsfor Action Modeling and Video Analysis , Eur. Conf. on Comp. Vis. 01-OCT-10, . : ,
08/13/2011	2.00	Alper Ayvaci, Stefano Soatto. Efficient model selection for detachable object detection, Proceedings of EMMCVPR. 24-JUL-11, . : ,
08/17/2011	4.00	Taehee Lee, Stefano Soatto. Learning and matching multiscale template descriptors for real- time detection, localization and tracking, Intl. Conf. on Comp. Vis. Patt. Recog.. 06-JAN-11, . : ,
08/26/2012	11.00	Michalis Raptis, Iasonas Kokkinos, Stefano Soatto. Discovering discriminative action parts from mid-level video representations, 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 15-JUN-12, Providence, RI, USA. : ,
08/26/2012	12.00	Michalis Raptis, Stefano Soatto. multiple instance filgering, neuro information processing systems (NIPS). 08-DEC-12, . : ,

TOTAL: **9**

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received		Paper
08/26/2012	8.00	Taehee Lee, Stefano Soatto. Video-based descriptors for object recognition, Image and Vision Computing (02 2011)
TOTAL:		**1**

Number of Manuscripts:

Books

Received		Book
08/13/2011	6.00	Stefano Soatto. Actionable information in vision, Berlin: Springer Verlag, (06 2011)
08/26/2012	7.00	Jingming Dong and Stefano Soatto. Visual Correspondence, the Lambert-Ambient Shape Space and the Automatic Design of Feature Descriptors, London: Springer, (08 2012)
TOTAL:		**2**

Received		Book Chapter
TOTAL:		

Patents Submitted

Patents Awarded

Awards

IEEE Fellow, 2013
Keynote Speaker: ICVSS2014, ICVSS2013, IEEE GlobalSIP 2014
Distinguished Seminars: Minnesota 2014, CVPR2014 Long-term detection and tracking, Mathematics and Image Analysis (Henri Poincare Institute, 2012)
Program Co-Chair: SIAM Imaging 2016, ICCV 2017
Area Chair: ICCV 2013, ACCV 2014
Associate Editor: SIAM Imaging, Intl. J. of Comp. Vision, Foundations and Trends in Graphics and VIsion

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Damek Davis	0.10	
Jingming Dong	0.10	
Georgios Georgiadis	0.20	
Nikolaos Karianakis	0.10	
Taihee Lee	0.10	
Daniel O'Connor	0.30	
Brian Taylor	0.10	
Konstantine Tsotsos	0.10	
Joachim Valente	0.10	
Zhao Yi	0.10	
FTE Equivalent:	**1.30**	
Total Number:	**10**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
Avinash Ravichandra	0.10
Ganesh Sundaramoorthi	0.10
FTE Equivalent:	**0.20**
Total Number:	**2**

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Stefano Soatto	0.33	
FTE Equivalent:	**0.33**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED	Discipline
Timothy Brightbill	0.20	Computer Science
FTE Equivalent:	**0.20**	
Total Number:	**1**	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 1.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 1.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 1.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:..... 0.00

Names of Personnel receiving masters degrees

NAME
Joachim Valente
Total Number: **1**

Names of personnel receiving PHDs

NAME
Taihee Lee
Daniel O'Connor
Zhao Yi
Total Number: **3**

Names of other research staff

NAME	PERCENT_SUPPORTED
Jason Meltzer	0.10
FTE Equivalent:	**0.10**
Total Number:	**1**

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See attachments

Technology Transfer

Visual Information Theory and Visual Representation
ARO 56765-CI
Final Report – September 30, 2013

Stefano Soatto
UCLA Vision Lab
University of California, Los Angeles

This project pursued the development of representations of visual data suitable for control and decision tasks. The fundamental premise is that traditional notions of information developed in support of communication engineering – where the task is reproduction of the source data, and nuisance factors can be easily characterized statistically – are unsuited to visual inference, where the task is decision or control, and the data formation process include *scaling* (that makes the continuous limit relevant) and *occlusion* (that makes control relevant).

Specifically, the task (or classes of task) inform what portion of the data is "informative" and what is "nuisance variability." One of the peculiarities of visual processing is that most of the complexity in visual data can be ascribed to nuisance factors that affect the data but are irrelevant to the task. This notion has been made precise in [16], preceding the commencement of this project, where it was shown that the *quotient* of the (infinite-dimensional) set of image modulo changes of *viewpoint and illumination is supported on a set of measure zero of the image domain.* So, for any task that requires invariance to viewpoint and illumination (such as object detection, localization, recognition, categorization), a zero-measure set contains as much "information" as the original data. In other words, information is a *thin set* of visual data, for decision and control tasks.

The development of a theory of information in support of decision and control tasks, specific to visual data, is a long-term goal that has been under development since 2007, and will continue for the foreseeable future. During the course of this project, significant progress has been registered in a number of areas critical to such a development, which is described below.

Below is a summary of the research accomplishments achieved during the period of performance, including progress since the last report dated July 31, 2012.

Actionable Information

We have continued progress on occlusion detection, previously reported in [1], that has been expanded and published in journal form in [2] and has been instrumental in the quantification of uncertainty in the presence of topological uncertainty due to occlusions. In fact, in the presence of visibility nuisances, the "innovation" (information increment) is precisely the complexity of the maximal invariant of the data relative to the nuisances *restricted to the discovered area* of the data domain. We have focused on images, but the results apply to any remote sensing modality including lidar, time-of-flight, etc.

While the formal definitions and the general principles cut across sensors, instantiating a model so that actual values can be computed can be quite challenging. Specifically, for passive remote sensing such as electro-optical, the *simplest model* that includes all the key ingredients (illumination, reflectance, shape, motion, occlusions, etc.) is the so-called *Lambert-Ambient* model, that is already rather complex to analyze. In [4] we have commenced an explicit analysis of this model for the purpose of characterization of Actionable Information. The quotient structure ("shape space") of this model has been fully characterized, and the results have been shown to yield explicit construction of visual feature descriptors that are – by construction – best suited for the specific task of object detection. This is another intermediate step in the design of a fully rational visual inference system that is driven by analytically sound principles as opposed to vague biological inspiration or trial-and-error practices.

One example of application of these principles is illustrated in [10], where we have been able to design descriptors that beat the state of the art in standard benchmark datasets, and in addition do so while significantly reducing the computational load, to the point where they can be implemented on mobile devices (smartphones) and operate in real-time (7-15FPS).

Inference

In the early part of this project, we have focused on – to a large extent solved – the problem of **occlusion detection**. Occlusion plays a fundamental role in the theory of Actionable Information, for the maximal viewpoint invariant in the un-occluded region represents the *innovation process*, or "information increment" that comes as a result of a control action. Therefore, it was very important that efficient and provably optimal occlusion detection algorithms be developed. Most of the existing literature on the topic was flawed in fundamental ways: First, most literature on motion estimation characterizes occluding boundaries as *motion discontinuities*. This is because such a literature is rooted in variational image processing, where the limit $dt \to 0$ is considered. When the inter-frame temporal interval goes to zero, clearly there are no occlusions. However, there is no motion either, so this literature completely neglected occlusions. More recent literature that tackles occlusion detection directly focuses on occluded regions as those where forward- and backward-motion are inconsistent. However, an occluded region is one that is visible in one image but not the next, and therefore the motion between these regions in the two images is not just inconsistent, *it is simply not defined*, it does not exist.

We have formalized the problem as a variational (infinite-dimensional) optimization problem, where one simultaneously tries to estimate the motion that maps one image onto the next (a diffeomorphism), together with the portion of the domain where such a map is defined. These are the co-visible regions, and their complement is the (multiply-connected) occluded region. While this appears to be a very difficult optimization problem, under the assumptions of Lambertian Reflection and constant (or slowly varying) ambient illumination, we have been able to show that the problem can be framed as a mixed ℓ^0/ℓ^2 optimization problem, whose re-weighted ℓ^1 relaxation yields a *convex optimization problem*. This has enabled us to exploit all the recent arsenal of efficient numerical schemes for convex optimization, including Augmented Lagrangian schemes such as Split-Bregman, as well as optimal first-order schemes due to Nesterov. We have made our source code publicly available, and many have been using it since. Furthermore, we have developed a GPU implementation of the algorithm, that exploits its parallelism, and achieved real-time implementation [5].

The importance of mobility for recognition is further emphasized in [9], that presents a real-time visual recognition system based on sparse representations that are guided by Actionable Information. Descriptors for static objects are built from multiple views, taken while the user moves around an object. While in this case the control loop is closed by the user, and the system assumes that the user acts rationally in collecting images of an objects that approximate a fair sample from the conditional distribution of a maximal invariant to nuisance factors, eventually we expect to design closed-loop robotic systems that perform such a "visual exploration" automatically.

Semantic Video Segmentation

Occlusion detection is critical to bootstrap the relation between structures in the *images* of a video, and the topology of the *scene*. In addition, one may be able to associate regions of images to "labels", in which case the notion of Detachable Object – bootstrapped from occlusion relations – provides a *semantic* segmentation of the video, where objects are represented as simply connected regions that back-project onto piecewise continuous surfaces in space that have *consistent labeling*. It also provides *relations* between objects, in the sense that the scheme does not just attach labels to object, but also determines whether there are multiple objects, and in what depth ordering they are presented relative to the viewer. The key results have been presented in [17], where the ideas have been tested on benchmark datasets.

Active Learning and Active Inference

In order to infer semantic segmentation of a video, a probability of detection should be provided at each pixel of each frame, for each possible category label, which is clearly a tall order. In the presence of millions of pixels per frame, hundreds of frames, and tens of objects, that would mean running billions of object detection algorithms just to process a few seconds of video.

Therefore, in [7] we have focused on *active inference*, where an information-gain criterion drives the decision as to *what frames* to process (when), *what pixels* within each frame (where), and *what detectors* to run (what), based on spatio-temporal regularity and context. While in principle one would not expect that running a subset of detectors on a subset of locations in a subset of frames could improve performance over the "paragon" setting, we have found that indeed full exploitation of spatio-temporal regularities and context enable *improving* the performance in the pixel-level detection of multiple classes in video, at a fraction of the computational cost. This may seem counter to the Data Processing Inequality, which is one of the driving principles of our work, with the conundrum addressed by the strong prior structure in natural and man-made scenes. In [7] we have performed test on benchmark video segmentation tasks with 19 categories.

Visual-inertial sensor fusion with application to humanoid motion estimation

While detachable object detection and semantic video segmentation provide *topological* relations between surfaces in the scene, and their relation to the viewer, full *geometric* modeling of the scene requires the estimation of the Euclidean motion of the sensor platform. This is a long-standing problem, where the UCLA Vision Lab has played a pioneering role, starting with the first observability analysis, to the first demonstration of real-time estimation of general structure-from-motion, to the first analysis of the observability of visual-inertial fusion.

During this project we have further pushed the envelope to include motion priors suitable for human (or humanoid) motions that are not well approximated by tightly tuned random walks [18].

Personnel supported on the project

See information uploaded on the ARO Report System.

Awards and honors

see list uploaded on ARO Report System

Teaching and Mentoring

In addition, material developed during the course of this project has been used in support of **teaching** and mentoring: Specifically, the fundamental building blocks and high-level concepts underlying Visual Information Theory [13] have been used to teach a quarter-long course at UCLA (CS269: Visual Information Theory), as well as to teach summer courses at the International Computer Vision Summer School (ICVSS) in Scicli, Italy, in 2011, 2012, 2013, and the US-Sino Summer School on Computer Vision, Pattern Recognition and Machine Learning in Chengdu, China in 2011. Some of the elements of the theory have been

collected into a book chapter for the lecture notes of ICVSS [14], while [13] is an evolving manuscript that eventually will be turned into a textbook.

Several students who have been **mentored** during the course of this project (some supported directly by this grant, some interacting indirectly with critical personnel but supported on other sources) have moved on to key positions in industry and academia, including the University of Oxford, U.K. (Prof. Andrea Vedaldi), Temple University (Dr. Haibin Ling), KAUST (Dr. Ganesh Sundaramoorthi), Google INC. (Dr. Alessandro Bissacco, Dr. Teresa Ko, Dr. Taehee Lee, Dr. Zhao Yi), Amazon INC. (Dr. Avinash Ravichandran), Honda Research (Dr. Alper Ayvaci), Comcast (Dr. Michalis Raptis).

Publications

The following references describe work that has been conducted during this project and acknowledge support by ARO: Year 1: [14, 3, 9, 15, 1, 5, 8, 12], Year 2: [2, 4, 10, 6, 11, 19], Year 3: [7, 17, 18].

Transitions

While complete **transitions** have not been accomplished during the period of performance, since the research focused on fundamental issues underlying the theoretical development of a theory of visual information, the research milestones accomplished enabled improvement of specific tasks that we envision will result in transitions in the near to mid-term future. These include visual-inertial sensor fusion [18], and semantic video segmentation [17].

References

[1] A. Ayvaci, M. Raptis, and S. Soatto. Optical flow and occlusion detection with convex optimization. In *Proc. of Neuro Information Processing Systems (NIPS)*, December 2010.

[2] A. Ayvaci, M. Raptis, and S. Soatto. Sparse occlusion detection with optical flow. *Intl. J. of Comp. Vision*, 97(3):322–338, 2012.

[3] A. Ayvaci and S. Soatto. Efficient model selection for detachable object detection. In *Proc. of EMMCVPR*, July 2011.

[4] J. Dong and S. Soatto. *Machine Learning for Computer Vision*, chapter Visual Correspondence, the Lambert-Ambient Shape Space and the Systematic Design of Feature Descriptors. R. Cipolla, S. Battiato, G.-M. Farinella (Eds), Springer Verlag, 2014.

[5] B. Fulkerson and S. Soatto. Really quick shift. In *Workshop on GPU For Computer Vision*, Crete, September 2010.

[6] E. Jones and S. Soatto. Visual-inertial navigation, localization and mapping: A scalable real-time large-scale approach. *Intl. J. of Robotics Res.*, Apr. 2011.

[7] V. Karasev, A. Ravichandran, and S. Soatto. Active frame, location, and detector selection for automated and manual video annotation. *Proc. of CVPR*, November 15, 2012, then April 15, then Nov. 8.

[8] T. Ko, S. Soatto, D. Estrin, and A. Cenedese. Cataloging birds in their natural habitat. In *Proc. of the ICPR Workshop on Vision in Natural Environments*, September 2010.

[9] T. Lee and S. Soatto. Learning and matching multiscale template descriptors for real-time detection, localization and tracking. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, pages 1457–1464, 2011.

[10] T. Lee and S. Soatto. Video-based descriptors for object recognition. *Image and Vision Computing*, 2011.

[11] M. Raptis, Y. Kokkinnos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *Proc. of the IEEE Intl. Conf. on Comp. Vis. and Patt. Recog.*, 2012.

[12] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *Proc. of ECCV*, September 2010.

[13] S. Soatto. *Steps Toward a Theory of Visual Information.* http://arxiv.org/abs/1110.2053, Technical Report UCLA-CSD100028, September 13, 2010 2010.

[14] S. Soatto. *Machine Learning for Computer Vision*, chapter Actionable Information in Vision. R. Cipolla, S. Battiato, G.-M. Farinella (Eds), Springer Verlag, 2011.

[15] S. Soatto and A. Chiuso. Controlled recognition bounds for scaling and occlusion channels. In *Proc. of the Data Compression Conference*, March 2011.

[16] G. Sundaramoorthi, P. Petersen, V. S. Varadarajan, and S. Soatto. On the set of images modulo viewpoint and contrast changes. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, June 2009.

[17] B. Taylor, A. Ayvaci, A. Ravichandran, and S. Soatto. Semantic video segmentation from occlusion relations via convex optimization. In *Proc. of EMMCVPR*, August 2013 2013.

[18] K. Tsotsos, A. Pretto, and S. Soatto. Visual-inertial ego-motion estimation for humanoid platforms. In *Proc. IEEE Intl. Conf. Humanoid Robots*, Dec. 2012.

[19] K. Wnuk and S. Soatto. Multiple instance filtering. In *Proc. of NIPS*, 2011.